



Gauss's, Cholesky's and Banachiewicz's Contributions to Least Squares

Gustavson, Fred G.; Wasniewski, Jerzy

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):
Gustavson, F. G., & Wasniewski, J. (2011). *Gauss's, Cholesky's and Banachiewicz's Contributions to Least Squares*. Technical University of Denmark. IMM-Technical Report-2011 No. 16

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Gauss's, Cholesky's and Banachiewicz's Contributions to Least Squares

Fred G. Gustavson¹ and Jerzy Waśniewski²

¹ IBM Watson Research, Emeritus and Umeå University
email: fg2935@hotmail.com

² Technical University of Denmark, Emeritus
email: jw@imm.dtu.dk

Abstract. This paper describes historically Gauss's contributions to the area of Least Squares. Also mentioned are Cholesky's and Banachiewicz's contributions to linear algebra. The material given is backup information to a Tutorial given at PPAM 2011 to honor Cholesky on the hundred anniversary of his algorithm.

The paper is also registered as a DTU/IMM¹-Technical-Report-2011-16.
Updated: August 18, 2011.

¹ DTU – Technical University of Denmark
IMM – Department of Informatics & Mathematical Modeling

1 Introduction

We are giving a Tutorial to honor both Cholesky and Banachiewicz at PPAM 2011 in Toruń, Poland. Recently, Grcar [14, 15] presented two historical papers on how common elimination became Gaussian Elimination. On pages 785 to 786 of [15] and on pages 15 to 23 [14] he discusses Gauss's contributions to the area of least squares; the latter [14] gives a more in depth treatment. Another view was given by Stewart in [24–26]. We follow Stewart's treatment here. [24] gives a modern and lucid description of Gauss's contributions. Hence we saw no reason to rephrase a lot of what Stewart had translated or to change his explanations thereof. Matrices were unknown in the early 1800's but it is clear from Gauss's work that he understood the concepts and theory that matrix usage would bring. So, one can make a strong case for giving Gauss credit for definitively laying down the early theory and practice of solving $Ax = b$ where A is an order n matrix; see Stewart's Afterword on pages 207-235 of [25]. This paper is meant to supplement the PPAM 2011 tutorial given by the two authors; the additional references not cited herein refer to the tutorial presentation. In Section 2 we overview some early history about least squares. Section 3 to 5 are devoted to the contributions of Gauss, Cholesky and Banachiewicz to least squares.

2 Overview of the early history of Least Squares

The principles of least squares arose from the problem of combining sets of overdetermined equations to form a square system that could be solved for its unknowns. Gauss kept a notebook in which he noted his important discoveries. In June 1798 he writes: "The problem of elimination resolved in such a way that nothing more can be desired". Stewart thinks this entry to be the first reference to Gaussian elimination. Also, see footnote *c* of Article 37, Section II of [6]. Here Gauss solves the problem of computing $Ax = b \pmod{m}$ where A is a matrix of integers. In the footnote Gauss says that "... a similar paralgoism exists for solving linear equations". At the end of 1801 Gauss had predicted, using his Least Square calculations, where the asteriod Ceres would be found! He was correct; this made his reputation! Gauss, who was generally slow to publish, began work in 1805 on his *Theoria Motus Corporum Coelestrium*, in which he described his techniques for computing orbits and gave his first probablistic justification on the principles of least squares; see [7]. He finished in 1806, but his publisher worried by German losses to Napoleon, insisted he translate the treatise into Latin. In the meantime, Legendre published and named the method of least squares in an appendix to his memoir of 1805.

In the *Theoria Motus*, Gauss had assumed the errors in the observations were normally distributed. In 1811, Laplace used his central limit theorem to give an essentially different justification of least squares. Laplace showed that the solutions of a combination of equations were asymptotically normal and from this concluded that the least squares combination would minimize the mean absolute

error in the solutions. Laplace's approach does not readily extend beyond two unknowns.

The final chapter occurred in the 1820's when Gauss [10, 11, 13] published two memoirs on least squares. The first, in two parts, contains yet another justification of least squares – Gauss's famous minimum variance theorem. These papers also contain some nice algorithms.

3 Matrix form of Gauss's Theory Motus paper of 1809

Theory Motus [7] is a long paper. We start Section 3 well into it at Section 182 of [7]. We use matrices which had not been invented and proceed as follows: Let $\mathbf{y} = \mathbf{A}\mathbf{b} + \mathbf{e}$ where \mathbf{A} is n by m , \mathbf{y} and \mathbf{e} are n vectors, \mathbf{b} is an m vector and $n > m$. The distribution of the errors e_i are assumed to be independent random variables with common distribution $\phi(e)$. Gauss introduces the function

$$\prod_{i=1}^n \phi(y_i - \mathbf{a}_i^T \mathbf{b}) \quad (1)$$

where \mathbf{a}_i^T is row i of \mathbf{A} . He uses a Bayesian argument with a uniform prior to show the value of \mathbf{b} that maximizes (1) is the most probable value of the unknowns. Reverend Thomas Bayes (1702-61), English mathematician and theologian, is responsible for his Theorem on conditional probability. We partly describe Bayesian inference: a parameter is assumed to have a *prior* distribution reflecting the experimenters belief in the state of nature. An experiment is performed, and the information from it, summarized by the likelihood, is combined with the prior distribution to provide a *posterior* distribution for the parameter.

Gauss assumes the distribution of the errors e_i to be normal; i.e., $\phi(e) \propto e^{-h^2 e^2}$. The parameter h gives the precision of \mathbf{y} . The function (1) now becomes proportional to

$$e^{-h^2 \Omega} \quad (2)$$

where

$$\Omega = \mathbf{e}^T \mathbf{e} = (\mathbf{y} - \mathbf{A}\mathbf{b})^T (\mathbf{y} - \mathbf{A}\mathbf{b})$$

is the residual sum of squares. The normal equations are derived by taking partial derivatives of Ω with respect to b . Gauss now turns to estimating the precision of the least square estimates. He integrates all but the last unknown out of (2). The precision can then be read off. However, to perform the required integrations $\Omega = \sum_{i=1}^n e_i^2$ where $e_i = y_i - (\mathbf{A}\mathbf{b})_i$ must be expressed in a special form, and the tool for arriving at the form is describing Gaussian Elimination implicitly!

The procedure as given by Gauss is the following. Let

$$u_1 = \frac{1}{2} \frac{\partial \Omega}{\partial b_1} \equiv \sum_{i=1}^m r_{1i} b_i - s_1 \quad (3)$$

and let

$$\Omega_1 = \Omega - \frac{u_1^2}{r_{11}} \quad (4)$$

Note that $r_{1j} = \sum_{k=1}^n a_{1k}a_{jk}$ and $s_1 = \sum_{k=1}^n a_{1k}y_k$. Then clearly $\partial\Omega/\partial b_1 = 0$ and so Ω_1 is independent of b_1 . We give another step to illustrate Gauss's general procedure. Set

$$u_2 = \frac{1}{2} \frac{\partial\Omega_1}{\partial b_2} \equiv \sum_{i=2}^m r_{2i}b_i - s_2$$

Then $\Omega_2 = \Omega_1 - u_2^2/r_{22}$ is independent of b_1 and b_2 . Continuing this way we arrive at the decomposition

$$\Omega = \sum_{k=1}^m \frac{u_k^2}{r_{kk}} + \rho$$

in which u_k is independent of b_1, \dots, b_{k-1} and ρ is a constant.

Gauss now considers the expression

$$e^{-h^2\Omega} \propto \exp(-h^2 \frac{u_1^2}{r_{11}}) \times \exp(-h^2 \frac{u_2^2}{r_{21}}) \dots \exp(-h^2 \frac{u_m^2}{r_{mm}}).$$

and integrates with respect to b_1 over the real line. Since the last $m-1$ factors in this expression are free of b_1 , they remain unchanged by integration. The first factor integrates to a constant. Thus Gauss is left with a distribution proportional to

$$e^{-h^2\Omega_1} \propto \exp(-h^2 \frac{u_2^2}{r_{21}}) \dots \exp(-h^2 \frac{u_m^2}{r_{mm}}).$$

which is free of b_1 . Continuing this process of integrating out the parameters b_i , Gauss finds the distribution of b_m is proportional to $\exp(-h^2 u_m^2/r_{mm})$ where $u_m = r_{mm}b_m - s_m$. Gauss concludes that the most probable value of b_m , obtained by setting $u_m = 0$, is $\hat{b}_m = s_m/r_{mm}$ and its precision is $h/\sqrt{r_{mm}}$.

Gauss now goes on to show that if you write the normal equations in the form

$$\mathbf{B}\mathbf{b} = \mathbf{c} \tag{5}$$

where $\mathbf{B} = \mathbf{A}^T\mathbf{A}$ and express \mathbf{b} as a function of \mathbf{c} in the form

$$\mathbf{b} = \mathbf{V}\mathbf{c}, \tag{6}$$

then $v_{mm} = 1/r_{mm}$. Finally, since the resulting expression for the precision clearly does not depend on the position of the unknown, Gauss concludes that the precision of *any* of the estimates \hat{b}_i is $h\sqrt{v_{ii}}$.

This is what Gauss does in *Theoria Motus*. He does *not* describe Gaussian elimination except implicitly! However, he gives a derivation of one of the most important results of linear regression theory! He also promises to describe his elimination procedure in another work. Gauss does point out that the normal equations can be solved by ordinary elimination (*eliminatio vulgaris*). Today, ordinary elimination is $\mathbf{L}\mathbf{U} = \mathbf{P}\mathbf{A}$ or perhaps Gauss Jordan elimination. An extension, which Gauss will later call general elimination (*eliminatio indefinita*), can be used to pass from the normal equations (5) to the inverse system (6).

3.1 The Notation that Gauss used

Few mathematicians in Gauss's time used subscripts or superscripts. Gauss wrote linear systems in the form

$$\begin{aligned} v &= ax + by + cz + etc. + l \\ v' &= a'x + b'y + c'z + etc. + l' \\ v'' &= a''x + b''y + c''z + etc. + l'' etc. \end{aligned}$$

Here $x, y, z, etc.$ are the unknowns we have been denoting by b_i and the v 's are the errors. This notation seems to be based on the way "one counts alphabetically" which is what subscripting does with integers. Gauss is quite good at it. Here is how he writes (slightly edited) the normal equations.

$$\begin{aligned} 0 &= [aa]x + [ab]y + [ac]z + etc. + [al] \\ 0 &= [ab]x + [bb]y + [bc]z + etc. + [bl] \\ 0 &= [ac]x + [bc]y + [cc]z + etc. + [cl] etc. \end{aligned}$$

Note the elegant way in which the notation $[ab]$ suggests a dot product (\mathbf{a}, \mathbf{b}) from the vectors of the a and b columns.

Gauss's notation for elimination is equally well considered. The following is from the *Supplementum* [13, 25] to the *Theoria Combinationis* (Article 13).

$$\begin{aligned} [bb, 1] &= [bb] - [ab]^2/[aa] \\ [bc, 1] &= [bc] - [ab][ac]/[aa] \\ [bd, 1] &= [bd] - [ad][ad]/[aa] \\ etc. \\ [cc, 2] &= [cc] - [ac]^2/[aa] - [bc, 1]^2/[bb, 1] \\ [cd, 2] &= [cd] - [ac][ad]/[aa] - [bc, 1][bd, 1]/[bb, 1] \\ etc. \\ [dd, 3] &= [dd] - [ad]^2/[aa] - [bd, 1]^2/[bb, 1] - [cd, 2]^2/[cc, 2] \\ etc. etc. \end{aligned}$$

Here as above, a pair of letters indicates the position in the normal equations. The appended numerals indicate the level of elimination. Gauss starts with $\mathbf{R} = \mathbf{A}^T \mathbf{A}$ and does a row by row $\mathbf{U}^T \mathbf{D} \mathbf{U} = \mathbf{R}$ factorization of the symmetric matrix \mathbf{R} . This is a preferred form for hand calculation, since one need only record an array of $m(m+1)/2$ numbers. Since \mathbf{R} is also positive definite one gets Cholesky's algorithm $\mathbf{V}^T \mathbf{V}$ where $\mathbf{V} = \sqrt{\mathbf{D}} \mathbf{U}$. Note however, that this work was done by Gauss in the 1820's.

3.2 The Gaussian Elimination Algorithm of the *Theoria Motus* ... paper

In 1810, in *Disquisitio de Elementis Palladis*, [8], Gauss gave the details of his algorithm for [7] of Section 3 and illustrated it with an example. The usual scalar formulas for the outer product form of Gaussian elimination of $\mathbf{R} = \mathbf{A}^T \mathbf{A}$ can

be obtained from Section 3 by starting with the second partial derivatives of Ω with respect to b . Specifically, set

$$r_{ij} = \frac{1}{2} \frac{\partial^2 \Omega}{\partial b_i \partial b_j}.$$

then we see from (4) that

$$\Omega_1 = \Omega - \frac{1}{r_{11}} \left(\frac{1}{2} \frac{\partial \Omega}{\partial b_1} \right)^2.$$

It follow that

$$r_{ij}^{(1)} \equiv \frac{1}{2} \frac{\partial^2 \Omega_1}{\partial b_i \partial b_j} = r_{ij} - r_{i1} r_{1j} / r_{11}.$$

In the expression on the right, we recognize a rank one update or the outer product form of Gaussian Elimination, in common practice today, on a matrix whose elements are r_{ij} . This is essentially the algorithm Gauss describes in *Disquisitio*, [7]. To complete the solution of the normal equations by Gaussian Elimination, note that since

$$\Omega = \sum_{k=1}^m \frac{u_k^2}{r_{kk}} + \rho$$

the function Ω assumes it minimum value ρ when

$$u_1 = u_2 = \dots = u_m = 0$$

Since

$$0 = u_m = r_{mm} b_m - s_m$$

is a linear equation involving only b_m , it can be solved immediately for b_m . Having determined b_m , one can solve for b_{m-1} from the equation

$$0 = u_{m-1} = r_{m-1,m-1} b_{m-1} + r_{m-1,m} b_m - s_{m-1}.$$

Continuing in this manner, we can determine estimates for all the unknowns b_i . This of course is nothing more than the back substitution phase of Gaussian elimination for a positive definite symmetric matrix.

The above description of the algorithm is incomplete, in the sense that it does not contain the forward elimination part of Gaussian elimination. This part comes from the formulas of Section 3, see (3) with $i = 1$, that give the constant parts s_i of the functions u_i . To see these connections it will be useful to express the algorithm in terms of matrices. The function Ω can be written in the form

$$\begin{aligned} \Omega &= (\mathbf{b}^T - 1) \begin{pmatrix} \mathbf{A}^T \mathbf{A} & \mathbf{A}^T \mathbf{y} \\ \mathbf{y}^T \mathbf{A} & \mathbf{y}^T \mathbf{y} \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ -1 \end{pmatrix} \\ &\equiv (\mathbf{b}^T - 1) \begin{pmatrix} \mathbf{B} & \mathbf{c} \\ \mathbf{c}^T & \eta \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ -1 \end{pmatrix} \end{aligned}$$

If we set

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ 0 & r_{22} & \dots & r_{2m} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & r_{mm} \end{pmatrix} \text{ and } \mathbf{s} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix},$$

where the r 's and s 's are from the definitions of the function u , see again (3) with $i = 1$, then it is easy to verify that

$$\begin{pmatrix} \mathbf{B} & \mathbf{c} \\ \mathbf{c}^T & \eta \end{pmatrix} = \begin{pmatrix} \mathbf{R}^T & 0 \\ \mathbf{s}^T & \rho \end{pmatrix} \begin{pmatrix} \mathbf{D}^{-1} & 0 \\ 0 & \rho^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{R} & \mathbf{s} \\ 0 & \rho \end{pmatrix},$$

where

$$\mathbf{D} = \text{diag}(r_{11}, r_{22}, \dots, r_{mm}).$$

Thus, Gaussian elimination, as practiced by Gauss, amounts to factoring the *augmented cross-product matrix* into the transpose of an upper triangular matrix, a diagonal matrix and an upper triangular matrix. It is common practice today to work with the augmented cross-product matrix.

The vector \mathbf{u} whose components are the functions u_i can be written in the form

$$\mathbf{u} = \mathbf{R}\mathbf{b} - \mathbf{s}.$$

The process sketched above of setting the u_i to zero and back-solving amounts to solving the triangular system

$$\mathbf{R}\mathbf{b} = \mathbf{s}.$$

3.3 The Computation of Variances

Writing in 1821, Gauss [9] summarized his and Laplace's justifications of least squares as follows:

- From the foregoing we find that the two justifications each leave something to be desired. The first depends entirely on the hypothetical form of the probability of the error; as soon as that form is rejected, the values of the unknowns produced by the method of least squares are no more the most probable values than is the arithmetic mean in the simplest case mentioned above. The second justification leaves us entirely in the dark about what to do when the number of observations is not large. In that case the method of least squares no longer has the status of a law ordained by the probability calculus and has only the simplicity of the operations it entails to recommend it.

In the *Pars Prior* of his memoir *Theoria Combinationis Observationum Erroribus Minimus Obnoxiae* [12], Gauss resolved the dilemma by introducing the notion of mean square error as a measure of variance and showing that among all linear combinations of the observations that produced exact estimates in the absence of error the least square estimates have the least mean square error.

In the *Pars Posterior of Theoria Combinationis* [11], Gauss addresses the problem of computing variances. He points out that his elimination method gives only the variance of the last unknown. Since (he continues) a general elimination to invert the normal equations is expensive, some calculators have adopted the practice of performing the elimination with another unknown placed last ². Gauss says that he will give a better way.

Gauss actually gives two solutions to the problem. In the first he shows that if one inverts the system $\mathbf{R}\mathbf{b} = \mathbf{s}$ to get $\mathbf{T}\mathbf{s} = \mathbf{b}$, then the matrix \mathbf{V} obtained by passing from (5) to (6) can be written

$$\mathbf{V} = \mathbf{T}\mathbf{D}\mathbf{T}^T.$$

Thus the diagonal elements of \mathbf{V} can be computed as weighted sum of squares of the row of \mathbf{T} . Gauss gives two algorithms for Computing \mathbf{T} , one of them particularly advantageous when only a few variances are to be computed.

The second method is a very general result for computing the variance of an arbitrary linear combination

$$t = \mathbf{g}^T \mathbf{b} + \kappa$$

of the unknowns \mathbf{b} . Specifically, if we pass from the variables \mathbf{b} to the variables \mathbf{u} , so that t assumes the form

$$t = \mathbf{h}^T \mathbf{u} + \hat{t},$$

then \hat{t} is the value of t at the least squares estimate of the unknowns ³, and its variance is proportional to

$$\mathbf{h}^T \mathbf{D} \mathbf{h}.$$

Moreover, \mathbf{h} may be obtained by solving the triangular system

$$\mathbf{R}^T \mathbf{h} = \mathbf{g}.$$

Thus Gauss reduces the problem of computing a variance to that of solving a triangular system.

A modern practice in numerical linear algebra is to compute a matrix decomposition and then use it in a variety of computations. The results of Gauss's elimination here serve as a computational platform from which both estimates and variances can be obtained.

4 Cholesky, 1875-1918 and his Algorithm

In [3] Brezinski gives a short history of the life of Andre Marie Cholesky. He was killed on a battle field of World War I which indeed was a tragedy! Cholesky's

² Laplace, for example, recommended a similar procedure in the first supplement to his paper *Théorie Analytique des Probabilités*.

³ It has been asserted by H. L. Seal in *Biometrika*, 54:1-24, 1967 that Gauss established that \hat{t} enjoyed the same minimum variance properties as the components of $\hat{\mathbf{b}}$. Although the result is true, Gauss never proved it.

algorithm was written down by him on December 2, 1910. We review it here as detailed in [3, 1, 14]. Nearly 100 years transpired before Cholesky's original manuscript was studied and reported by Brezinski. However, Cholesky's colleague was Commander Benoit and he, in [1], in 1924 published a description of Cholesky's algorithm. Cholesky's algorithm was used for geodesy applications and it appears to have been developed for pinwheel type calculating machines; e.g., Cholesky used the Dactyle machine. Cholesky calculated Gauss's angle adjustments; these were formulated as an underdetermined least squares problem. Cholesky did not use matrices even though matrices were known before he was born. So, his invention is perhaps better understood in algebraic notation. However, we shall also carry along a matrix description to help modern readers. We will therefore use a slightly different notation than either Cholesky and Benoit used but also use their scalar notation so as to be compatible with their descriptions. We write the so called condition equations as

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n + K_1 &= 0 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n + K_2 &= 0 \\ \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n + K_m &= 0 \end{aligned} \tag{7}$$

The condition equations (7) have the form $\mathbf{Ax} + \mathbf{K} = \mathbf{0}$ where \mathbf{A} is a m by n matrix, \mathbf{x} is a n vector, \mathbf{K} is a m vector and $m < n$.

$$\begin{aligned} b_{11}y_1 + b_{12}y_2 + \dots + b_{1m}y_m + K_1 &= 0 \\ b_{21}y_1 + b_{22}y_2 + \dots + b_{2m}y_m + K_2 &= 0 \\ \vdots & \\ b_{m1}y_1 + b_{m2}y_2 + \dots + b_{mm}y_m + K_m &= 0 \end{aligned} \tag{8}$$

The above normal equations (8) have the form $\mathbf{By} + \mathbf{K} = \mathbf{0}$ where $\mathbf{B} = \mathbf{AA}^T$ is a m by m positive definite symmetric matrix, and $\mathbf{y} = \mathbf{Ax}$ is a m vector. Thus $b_{ij} = b_{ji}$ and b_{ij} is the dot product (sum of products of coefficients in the i^{th} and j^{th} conditions of (7).

Cholesky's remarkable insight was that, because many underdetermined systems share the same normal equations, for any normal equations there may be some condition equations that can be directly solved more easily; see page 70 of [1]. He found his alternate equations in the convenient triangular form (introducing new unknowns, z in place of x , and new coefficients, c_{ij} in place of b_{ij} . The b 's come from matrix \mathbf{B} of equations (8) above and the c 's come from matrix \mathbf{C} in equations (9) below.

$$\begin{aligned} c_{11}z_1 + & & & + K_1 &= 0 \\ c_{21}z_1 + c_{22}z_2 & & & + K_2 &= 0 \\ \vdots & \vdots & \ddots & \vdots & \\ c_{m1}z_1 + c_{m2}z_2 + \dots + c_{mm}z_m + K_m &= 0 \end{aligned} \tag{9}$$

To get equations (9) above we let $\mathbf{y} = \mathbf{Cz}$ where \mathbf{C} is a lower triangular matrix. Thus we find the matrix identity of $\mathbf{CC}^T = \mathbf{B}$. What this says is that the positive

definite symmetric matrix \mathbf{B} has the provable \mathbf{LU} matrix decomposition \mathbf{LL}^T . Gauss gave this decomposition as $U^T DU = B$ in [13]. Cholesky discovered that the coefficients in equation (9) are given by straightforward formulas (10); see page 72 of [1].

$$\begin{aligned} c_{ii} &= \sqrt{b_{ii} - \sum_{k=1}^{i-1} c_{ik}^2} & 1 \leq i \leq m \\ c_{ji} &= (b_{ji} - \sum_{k=1}^{i-1} c_{jk}c_{ik})/c_{ii} & i < j \leq m \end{aligned} \quad (10)$$

The solution \mathbf{y} of the normal equations expresses the solution of the condition equations as a linear combination of the (transposed) coefficients in the condition equations. For Cholesky, these combinations become a system of equations to be solved for \mathbf{y} ; see Eq. 7 of [1].

$$\begin{aligned} z_1 &= c_{11}y_1 + c_{21}y_2 + \dots + c_{m1}y_m \\ z_2 &= c_{22}y_2 + \dots + c_{m2}y_m \\ &\vdots \\ z_m &= c_{mm}y_m \end{aligned} \quad (11)$$

Because the original condition equations (7) and Cholesky's equations (9) have the same normal equations (8), the quantities y_i are the same for both problems. Once obtained, the y_i can be used to evaluate the x_i from the transpose of the original coefficients and thereby solve the original condition equations (7). Thus Cholesky's method was first to form his new coefficients c_{ij} using his formulas (10) and simultaneously solve $\mathbf{Cz} + \mathbf{K} = \mathbf{0}$ (equation (9)) by forward substitution for z_i . Next he solves $\mathbf{z} = \mathbf{C}^T \mathbf{y}$ (equation (11)) by backward substitution for y_i , and finally he evaluates x_i using equations (12) below:

$$\begin{aligned} x_1 &= a_{11}y_1 + a_{21}y_2 + \dots + a_{m1}y_m \\ x_2 &= a_{12}y_1 + a_{22}y_2 + \dots + a_{m2}y_m \\ &\vdots \\ x_n &= a_{1n}y_1 + a_{2n}y_2 + \dots + a_{mn}y_m \end{aligned} \quad (12)$$

Above we have followed Benoit's [1] paper as described by Grcar [14]. We have changed their notation as follows. For our equation (7) Benoit's equation (1) used coefficients a_i , b_i , l_i for $i = 1$ to n for rows one, two and m . For our equation (8) Benoit's equation (5) used coefficients a_i^j and vector λ for our vector \mathbf{y} . For our transformation (9), $\mathbf{z} = \mathbf{C}\mathbf{y}$, Benoit's transformation was $\mathbf{y} = \beta\lambda$.

4.1 Some Details of Cholesky's Original Algorithm

His algorithm was given exactly like the one Benoit gave. He gave explicit details about equations (7-12) so that soldiers (they had little mathematical background) could easily follow and perform its steps using the Dactyle mechanical calculating machine. In particular, he gave his soldiers a method for computing a square root: $x_{i+1} = (x_i + a/x_i)/2$ for \sqrt{a} . This formula can be obtained by the Newton Raphson general formula; however, it is due to Hero of Alexandria.

5 Cracovians and Tadeusz Banachiewicz, 1882-1954

Tadeusz was born in Warsaw (Russia occupied Poland). We fast forward his life until 1919 when Poland regained her independence. Banachiewicz then moved to Cracow (Poland), becoming a professor at the Jagiellonian University and the director of Cracow Observatory. In 1925, he invented a theory of "cracovians" (a special kind of matrix algebra) which brought him international recognition. Cracovians introduced the idea of using \mathbf{A}^T , and multiplying the columns of \mathbf{A}^T by the column vector \mathbf{x} to form the matrix vector product $\mathbf{y} = \mathbf{Ax}$. Thus Cracovians adopted a column-row convention for designating individual elements as opposed to the standard row-column convention of matrix analysis. This made manual multiplication easier, as one needed to follow two parallel columns (instead of a vertical column and a horizontal row in the matrix notation.) It also sped up computer calculations, because both factors' elements were used in a similar order, which was more compatible with the sequential access memory in computers of those times – mostly magnetic tape memory and drum memory. Use of Cracovians in astronomy faded as computers with bigger random access memory came into general use.

Using the cracovians Banachiewicz discovered general formulae of spherical polygonometry, and like Cholesky he considerably simplified the algorithm of the least squares method and the practice of finding solutions the systems of linear equations. The cracovian calculus has found numerous applications in spherical astronomy, celestial mechanics, determining orbits, and geodesy. The first orbit of Pluto was determined at the Cracow Observatory.

References

1. Benoit. Note sur une methode de resolution des equations normales provenant de l'application de la methode des moindres carres a un systeme d'equations lineaires en nombre inferieur a celui des inconnues. – Application de la methode a la resolution d'un systeme defini d'equations lineaires (Procédé du Commandant Cholesky). Bulletin geodesique (2), 67-77. 1924. The author is identified only as Commandant Benoit.
2. Carl. B. Boyer, revised by U. C. Merzbach. Determinants, Notation, and Imaginary Numbers. *A History of Mathematics*, book, 1968, 1991, John Wiley and Sons, pp. 405-407.
3. Brezinski, C. The life and work of André Cholesky. *Numerical Algorithms* Vol. 43 No. 3, pp. 279-288, 2006.
4. J. J. Dongarra, F. G. Gustavson, A. Karp. Implementing Linear Algebra Algorithms for Dense Matrices on a Vector Pipeline Machine. *SIAM Review*, Vol. 26, No. 1, Jan. 1984, pp. 91,112.
5. V. N. Faddeeva Chapter II: Systems of Linear Equations. *Computational Methods of Linear Algebra*, book, 1950, 1959, Dover Publications, Inc, NY, pp. 63-146.
6. C. F. Gauss Section II: Congruences of the First Degree; Art. 37, Footnote c, pp. 17-20. Section V: Forms and Indeterminate Equations of the Second Degree *Disquisitiones Arithmeticae*, book, 1801, 1966, Yale University Press, New Haven and London, 472 pages.

7. C. F. Gauss Theoria Motus Corporum Coelestrium in Sectionibus Conics Solem Ambientium. Perthes and Besser, Hamburg, 1809
8. C. F. Gauss Disquistitio de elementis ellipticis Palladis. *Commentatines societatis regiae scientarium Gottingensis recentiores* 1,1810
9. C. F. Gauss Anzeige: Theoria combinationis observationum erroribus minimis obnoxiae: Pars prior *Göttingische gelehrte Anzeigen*, 33:312:327, 1821
10. C. F. Gauss Anzeige: Theoria combinationis observationum erroribus minimis obnoxiae: Pars posterior *Göttingische gelehrte Anzeigen*, 32:313:318, 1823
11. C. F. Gauss Theoria combinationis observationum erroribus minimis obnoxiae: Pars posterior *Commentatines societatis regiae scientarium Gottingensis recentiores*, 5, 1823
12. C. F. Gauss Theoria combinationis observationum erroribus minimis obnoxiae: Pars prior *Commentatines societatis regiae scientarium Gottingensis recentiores*, 5, 1823
13. C. F. Gauss Supplementum theoria combinationis observationum erroribus minimis obnoxiae. *Commentatines societatis regiae scientarium Gottingensis recentiores*, 6, 1828
14. Joseph F. Grcar "How ordinary elimination became Gaussian elimination". *Historia Mathematica*, 38(2):163D218,2011a, 56 pages, 202 references
15. J. F. Grcar. Mathematicians of Gaussian Elimination. *Notices of the AMS* Vol. 58, No. 6, pp. 782-792, June-July, 2011.
16. F. G. Gustavson. Recursion Leads to Automatic Variable Blocking for Dense Linear-Algebra Algorithms. *IBM Journal of Research and Development*, Vol. 41, No. 6, Nov. 1997, pp. 737-755.
17. F. G. Gustavson High Performance Linear Algebra Algorithms using New Generalized Data Structures for Matrices. *IBM Journal of Research and Development*, Vol. 47, No. 1, Jan. 2003, pp. 31,55.
18. F. G. Gustavson, J. Waśniewski, J. J. Dongarra, J. Herrero, J. Langou. Level-3 Cholesky Factorization Routines as Part of Many Cholesky Algorithms. *Transactions on Mathematical Software*, in 2nd revision. also, LAWN No. 249, 23 pages, June, 2011.
19. Alston S. Householder. Chapter 5: Direct Methods of Inversion. *The Theory of Matrices in Numerical Analysis*, book, 1964, 1975, Dover Publications, Mineola, NY, pp. 122-146.
20. Graham Flegg. Chapter 6: Unknown Numbers. *NUMBERS: Their History and Meaning*, book, 1983, 2002, Dover Publications, Mineola, NY, pp. 203-212.
21. Joseph Louis Lagrange. Lecture 3: On Algebra, Particularly the Resolution of Equations of the Third and Fourth Degree. *Lectures on Elementary Mathematics*, book, 1795, 1898, 2008 Dover Publications, Mineola, NY, pp. 54-95.
22. C. D. Meyer. Matrix Analysis and Applied Linear Algebra, Book. *SIAM* Philadelphia, 2001.
23. David E. Smith Leibniz On Determinants. *A Source Book in Mathematics*, book, 1929, 1959, Dover Publications, Mineola, NY, pp. 267-270.
24. G. W. Stewart. Gauss, statistics, and Gaussian Elimination. Institute for Advanced Computer Studies TR-94-78 Dept. of Computer Science TR-3307, U. of Maryland, 8-94 13 pages.
25. G. W. Stewart. Theory of the Combination of Observations Least Subject to Errors it Classics in Applied Mathematics book, 1998, SIAM, Philadelphia, 1998.
26. G. W. Stewart. Matrix Algorithms Volume 1: Basic Decompositions, Book. it SIAM Philadelphia, 1998.

27. Ian Stewart. Chapter 15: Linear Algebra. *Concepts of Modern Mathematics*, book, 1995, Dover Publications, NY, pp. 215-228.
28. Gilbert Strang The Fundamental Theorem of Linear Algebra *The American Mathematical Monthly*, Vol. 100, No. 9. (Nov., 1993), pp. 848-855.